F. MASARANI AND S. S. GOKTURK

ON THE EXISTENCE OF FAIR MATCHING ALGORITHMS

ABSTRACT. We analyze the Gale–Shapley matching problem within the context of Rawlsian justice. Defining a fair matching algorithm by a set of 4 axioms (Gender Indifference, Peer Indifference, Maximin Optimality, and Stability), we show that not all preference profiles admit a fair matching algorithm, the reason being that even this set of minimal axioms is too strong in a sense. Because of conflict between Stability and Maximin Optimality, even the algorithm which generates the mutual agreement match, paradoxically, has no chance to be fair.

We then relax the definition of fairness (by giving preference to Stability over Maximin Optimality) and again find that some preference profiles admit a fair matching algorithm, while others still do not, but the mutual agreement algorithm now is fair under this definition.

The paper then develops a test, which determines, for a given preference profile, whether a fair algorithm exists or not.

Keywords: combinatorial optimization, theory of algorithms, matching, fairness, axiomatic justice, marriage, assignment markets, job matching.

INTRODUCTION

The problem of matching the members of one social group to members of another, such as men to women, students to colleges, employees to employers, etc. has generated considerable literature since it was first studied by Gale and Shapley (1962). However, the matching algorithms one encounters in this literature, such as Gale and Shapley (1962), Knuth (1976), Roth (1985 a, b), among others, all favor one group over another; and there has not been any analysis of the matching problem within the context of the theory of justice so far.

The purpose of this paper is to carry out such as analysis within the framework introduced by Gale and Shapley (1962). We prove in the course of this analysis that often (the circumstances being made precise in the text) it is impossible to construct a 'fair' matching algorithm, fairness being defined in the sense of Rawlsian justice (Rawls, 1971). Our

Theory and Decision **26** (1989) 305–322.

approach is in the same spirit as the original non-mathematical treatment of Rawls (1971) and its later mathematical analysis in Masarani and Gokturk (1986), in that all the parties who have gathered at the 'original position' behind a 'veil of ignorance' to decide on the rules that will govern their lives thereafter, including their pairwise matchings, will agree upon a set of axioms of justice.

A similar definition of fairness is also adopted by Rochford (1984), who shows that the outcome of an assignment game in her transferable utility model, which is quite different from those within the Gale–Shapley framework, is in the nucleolus and hence 'the result of an arbitrator's desire to minimize the dissatisfaction of the most dissatisfied coalition (in the game) (Rochford, 1984, p. 279).'

1. STATEMENT OF THE PROBLEM

Suppose that there are two groups of agents, which we will designate by X and Y, with each group consisting of n members. Suppose that each member of either group needs to be matched with exactly one member of the other group. Each agent has preferences by which he can rank the members of the opposite group from 1 to n, with 1 corresponding to the most desired member and n to the least desired one. The happiness of an agent with a match can be ascertained by how close to 1 he ranks the agent he is matched with.

Because of conflict in the preferences of the agents, or in order to improve the efficiency of the matching process, both groups agree to give their preferences to an entity and to empower that entity to match them, with the provision that the decisions of the entity will not be contested. In such a case we would like to pose and answer the following two questions: (1) What will constitute a fair match, and (2) will there always be a fair match, irrespective the structure of the preferences of the agents?

It is easy to see that this matching problem has many applications, some of which are already part of the literature. We list some of them by specifying the group X, the group Y, and nature of the match, in that order: (1) men, women, marriage, (2) students, universities, admission, (3) medical interns, hospitals, admission, (4) applicants for a certain job, firms which have openings for such a job, hiring, (5) buyers of co-ops, co-op boards, purchasing a co-op, assuming that all co-ops are identically

priced and all buyers can afford the price (see the references for some such applications).

Obviously, some of the above applications do not satisfy all our assumptions. For example, in (2), (3), and (4) the universities, hospitals, or firms may admit or hire more than one student, intern, or employee, respectively. However, we will show how our results do eventually apply to such cases with minor adjustments.

As we proceed to develop the model and derive the results of the paper, we will keep on referring to the matching of men and women. However, the words 'men' and 'women' are to be understood as generic names and not necessarily as biological men and women. Thus, when our results are being applied to, say, the problem of college admissions, the men could be interpreted as the students and the women as the universities. Similarly, we will interpret the words 'marriage' and 'divorce' in a more general sense than their usual meaning. Thus, the hiring of an employee by a firm is a 'marriage,' while the firing (or the quitting) of an employee a 'divorce.'

We now show how our model can be transformed to apply to situations where the number of individuals in X is not equal to the number of individuals in Y, or where a member of one of the two groups, say Y, needs to be matched with more than one member of the group X, with the use of a clever device (see Rochford, 1984). To be specific, we consider the case where Y are universities and X are students. Obviously, there are generally more students than universities and, while each student needs to be matched to only one university, each university needs to be matched to several students, whose number we will refer to as the quota of the university. We show how to cast this matching problem in stages, such that each stage satisfies the assumptions of our model.

We introduce in the first stage dummy universities, so that the total number of universities is equal to the number of students and let the ranking of the dummy universities by the students be the same for all the students and be at a higher level than the ranking of the real universities, while the ranking of the students by the dummy universities is chosen arbitrarily. When a student is matched to a dummy university in any one of the stages, we will not remove such a student from the set of eligible students, while, if a student is matched to a real university, we will remove that student from the set of eligible students, but keep the university, unless the university has already filled its quota.

2. TERMINOLOGY

We introduce in this section some definitions, which will enable us to state the matching problem formally.

A preference function on a set X is a function $f: \mathcal{P}(X) \to X$, where $\mathcal{P}(X)$ is the power set of X, which satisfies the following two axioms:

AXIOM 1: $\forall A \subseteq X: f(A) \varepsilon A$.

AXIOM 2: $R \supseteq S$ *and* $f(R) \varepsilon S \Rightarrow f(S) = f(R)$.

Let $\Pi(X)$ denote the set of preference functions on X.

Let X be a set of n men and Y a set of n women. Let $\Pi(X, Y) = \Pi^n(Y) \times \Pi^n(X)$. Any element of $\Pi(X, Y)$ is a $2n$-vector. Its first n components are preference functions on Y, which we will call the *preferences of men*, while the last n components are preference functions on X, which we will call the *preferences of women*.

The set $\Pi(X, Y)$ will be called the *set of preferences* and each one of its elements will be called a *preference profile*.

It is easy to see that a preference function can list the elements of X in a descending order of preference as follows: $f(X)$ is the most preferred. $f(X - f(X))$ is the second element of X in the order of preference. Continuing this process inductively, we obtain a sequence $x_1, x_2, \ldots, x_n$, which lists the elements of X, such as $x_1 = f(X)$, while $x_k = f(X_{k-1} - f(X_{k-1}))$, $k = 2, \ldots, n$, where $X_{k-1} = \{x_1, x_2, \ldots, x_{k-1}\}$. For any element $x_k \varepsilon X$, we will refer to k as the *rank* of x_k under the preference function f and to x_k the *kth choice* of f.

A *match* M of X and Y is a subset of $X \times Y$, such that $P_1(M) = X$ and $P_2(M) = Y$, where $P_1: X \times Y \to X$ and $P_2: X \times Y \to Y$ are the coordinate projections. The set $\mathcal{M}$ of all matches of X and Y will be called the *match space* of X and Y.

The *achievement profile* is a function $\alpha: \Pi(X, Y) \times \mathcal{M} \to \mathbb{Z}_+^{2n}$, where $\mathbb{Z}_+$ is the set of positive integers, defined as follows: $\alpha(p, M) = (\alpha_1(p, M), \alpha_2(p, M))$, where $p \varepsilon \Pi(X, Y)$, $M \varepsilon \mathcal{M}$, and $\alpha_1(p, M)$ is the n-vector whose components list the rank the men assign their spouses in the match M using their preference functions, while $\alpha_2(p, M)$ is the corresponding n-vector for the women.

A *matching algorithm* μ on Λ is a function $\mu: \Lambda \to \mathcal{M}$, where $\Lambda \subseteq \Pi(X, Y)$. For $p \in \Lambda$, $\mu(p)$ will be referred to as the *match generated by* μ.

3. CYCLIC PREFERENCES AND THE MUTUAL AGREEMENT MATCH

The matching of X and Y can be carried out for certain preference profiles by simply letting each man and woman choose freely his or her spouse. To see how this can happen, suppose that man x considers woman y his number one choice, and vice versa. Then, if man x were asked to propose to a woman, he certainly would choose woman y, and woman y would accept his proposal. Thus, x is matched to y, and they drop out from their respective sets of eligible individuals. If this process can be repeated until all the men and women are matched, then we will call such a match the *mutual agreement match.*

It is easy to see that a mutual agreement match is not possible for all preference profiles, for suppose that man x_1 considers woman y_1 his number one choice, who considers man x_2 her number one choice, who considers woman y_2 his number one choice, who considers man x_1 her number one choice. Then, it will not be possible to ever match x_1, x_2, y_1, and y_2, if matching is to be done subject to their mutual agreement, since each one of these individuals is waiting for the response of another to his (resp. her) marriage proposal (assuming that men and women are allowed to propose), who in turn is waiting the response of another, thus all ending up in a vicious circle. We will call the preference profile in such a situation *cyclic.* If a mutual agreement match can be carried out, we will call the preference profile *acyclic.*

4. FAIR ALGORITHMS AND JUSTICE PRINCIPLES

In what follows, we introduce the concept of a fair matching algorithm through a set of four axioms that such an algorithm must satisfy. In order that there be no ambiguity in stating some of the axioms, we will first state the definition of a matching algorithm in the language borrowed from the area of data processing.

We will think of a matching algorithm as a black box process with an input and an output. The input of the black box consists of two files which we will label as File 1 and File 2. Each file has n records. Each record

consists of a key, which identifies the individual occupying that record (such as the social security number) and a list of the records in the other file, which corresponds to the preferences of that individual. The preliminary output of the black box is a list of n ordered pairs, which matches each record in File 1 to a unique record in File 2. The final output of the black box consists of a list of n ordered pairs of keys, which corresponds to the match generated by the algorithm.

We now are in a position to list the axioms, which define a fair algorithm. Let $\mu: \Lambda \to \mathcal{M}$ be a matching algorithm. Then, μ will be called a *fair matching algorithm* over Λ, if it satisfies the following four axioms:

AXIOM 1. *Gender Indifference*: *For every $p \in \Lambda$, the final output μ is the same, whether the men (resp. women) are occupying File* 1 *or File* 2.

AXIOM 2. *Peer Indifference*: *For every $p \in \Lambda$, the final output of μ is the same, irrespective the order in which members of the same sex occupy the records of their respective files of the input.*

We need the following terminology for the statement of the next axiom. Let $p \in \Pi(X, Y)$ and $M \in \mathcal{M}$. The *floor* of the match M with respect to p is the positive integer $\ell(p, M)$ which is a largest component of the vector $\alpha(p, M)$. The *pedestrians* of the match M with respect to p are the individuals (men and women), who are matched by M to individuals whom they rank as their $\ell(p, M)$ choice.

An individual is *happier* in a match M_2 than he (resp. she) is in match M_1, if that individual prefers his (resp. her) spouse in M_2 to his (resp. her) spouse in M_1. Thus, if the rank of the spouse of this individual can be read from the ith component of the preference profile vector, then the individual is happier with the match M_2 than M_1, if and only if the ith component of $\alpha(p, M_2)$ is less than the ith component of $\alpha(p, M_1)$.

AXIOM 3. *Maximin Optimality*: *Let $p \in \Lambda$ be given and let $\mu(p) = M_0$. Then, there exists no match $M \in \mathcal{M}$, such that the pedestrians of M are happier than the pedestrians of M_0.*

We need the following definition for the next axiom. Let $p \in \Pi(X, Y)$ and $M \in \mathcal{M}$. We will say that M is *divorce-inclined* with respect to p, if there

exists at least one man and one woman, who prefer each other to their current spouses in the match M. If M is not divorce-inclined, we will call *it divorce-proof* or, to use the standard terminology of the literature, *stable*.

AXIOM 4. *Stability*: *For every* $p \in \Lambda$, $\mu(p)$ *is a stable match.*

We consider Axiom 4 a requirement of fairness, because, if a matching algorithm were permitted to generate unstable matches, then divorce has to be outlawed in order to prevent the individuals who are inclined to divorce from breaking their marriages. This prohibition of divorce, in our opinion, is unfair. In fact, the necessity of Axiom 4 is not reduced, even if our results are to be applied to biological men and women and 'marriage' and 'divorce' are to be interpreted in their usual meaning by a social institution, such as a government or a church, which is against divorce, for by requiring the fair algorithm to generate stable matches, we would insure that the temptation of divorce will not arise.

The axioms we postulated for the fairness of an algorithm obey Rawls' idea of choice under 'a veil of ignorance' (Rawls, 1971) in the following sense. If all the men and women were completely ignorant of their preferences and the preferences of others for them at the stage when they are going to sign a social contract with the matching entity, in which they and the matching entity agree on the general principles (axioms) by which the entity will match them, then all the individuals will insist that the matching entity employ Axioms 1 and 2, since each individual does not know a priori whether he (resp. she) will be listed in File 1 or File 2 of the matching algorithm or in what order he (resp. she) will be listed.

All the individuals will insist on Axiom 3, because, since they do not know as yet their preferences or the preferences of others, each individual is contemplating the possibility of ending up as a pedestrian in the match generated by the entity.

Axiom 4 is required, because all individuals demand the right to divorce at will and thus, if the matching algorithm is to have a chance to accomplish its goal, the matching entity has to constrain itself to producing stable matches.

A *justice principle* j over a subset Λ of preference profiles is a function j on Λ, which maps each preference profile into a partial order on the

set $\mathcal{M}$ and satisfies a set of four axioms to be listed below. We will denote the set of partial orders on $\mathcal{M}$ by $\mathcal{O}(\mathcal{M})$. Thus, in symbols, $j: \Lambda \to \mathcal{O}(\mathcal{M})$.

The partial order on $\mathcal{M}$, which is the image of a preference profile $p \in \Lambda$ under j, will be denoted by $\precsim j(p)$. The subscript $j(p)$ will be dropped, whenever the underlying p is understood from the context. For M_1 and M_2 in $\mathcal{M}$, the relationship $M_1 \precsim j(p) M_2$ is to be read as '*M_1 is not more fair than M_2*.' For technical reasons, we cannot read $M_1 \precsim j(p) M_2$ as 'M_2 is more fair than M_1,' since in that case the identity axiom of a partial order is not satisfied. What follows is a list of the axioms that the function j must satisfy in order to qualify as a justice principle.

AXIOM 1. *Non-chauvinism*: *For every $p \in \Lambda$ and matches $M_1, M_2 \in \mathcal{M}$, if $\alpha_2(p, M_1)$ can be obtained by permuting the components of $\alpha_1(p, M_2)$ and similarly for $\alpha_1(p, M_1)$ and $\alpha_2(p, M_2)$, then neither $M_1 \precsim j(p) M_2$ is true, nor $M_2 \precsim j(p) M_1$.*

AXIOM 2. *Peer Egalitarianism*: *For every $p \in \Lambda$ and matches $M_1, M_2 \in \mathcal{M}$, if $\alpha_1(p, M_1)$ can be obtained by permuting the components of $\alpha_1(p, M_2)$ and similarly for $\alpha_2(p, M_1)$ and $\alpha_2(p, M_2)$, then neither $M_1 \precsim j(p) M_2$ is true, nor $M_2 \precsim j(p) M_1$.*

AXIOM 3. *Rawls' Criterion*: *For every $p \in \Lambda$ and matches $M_1, M_2 \in \mathcal{M}$, if the floor of M_1 with respect to p is less than the floor of M_2 with respect to p, then $M_2 \precsim j(p) M_1$.*

AXIOM 4. *Majority Rule*: *For every $p \in \Lambda$ and matches $M_1, M_2 \in \mathcal{M}$, if the floor of M_1 with respect to p is equal to the floor of M_2 with respect to p and if a majority of individuals (men or women) are happier in M_1 than M_2, then $M_2 \precsim j(p) M_1$. If no such majority exists and $M_1 \neq M_2$, then neither $M_2 \precsim j(p) M_1$ is true, nor $M_1 \precsim j(p) M_2$.*

Given $p \in \Lambda$ and matches $M_1, M_2 \in \mathcal{M}$, it is easy to see that one of the above axioms can be invoked to determine whether $M_1 \precsim M_2$, $M_2 \precsim M_1$, or neither of the above, for any justice principle j. Furthermore, the axioms define a unique partial order on $\mathcal{M}$, for every $p \in \Pi(X, Y)$. Thus, we obtain the following

THEOREM 4.1. *Given any* $\Lambda \subseteq \Pi(X, Y)$, *then there exists a justice principle* j *on* Λ *and it is unique.*

Because of Theorem 4.1, we will refer to a justice principle as *the* justice principle, give it the title of *Rawls' justice principle*, and always denote it by the symbol j.

The reader may have noticed that there is a one-to-one correspondence between the Axioms 1 through 3 of a fair matching algorithm and the axioms of Rawls' justice principle and, thus, may have suspected the validity of the following

THEOREM 4.2. *Given* $p \in \Pi(X, Y)$ *and a fair algorithm* μ *over* p, *then* $\mu(p)$ *is a maximal element of* $j(p)$.

Proof. Let $\mu(p) = M_0$ and M be any element of $\mathcal{M}$. We prove that $M_0 \precsim j(p)\, M$ implies $M = M_0$. Assume $M \neq M_0$. Note that Axiom 3 of a fair algorithm and Axioms 3 and 4 of Rawls' principle of justice imply that the floor of M must be equal to the floor of M_0 and a majority of individuals is happier in M than in M_0. Then, at least one individual in such a majority has to be matched in M with another individual of the majority or a majority will not exist, since the vote of each individual i favoring M will be nullified by the vote of the individual with whom i is matched in M. But this implies that M_0 is unstable, which contradicts Axiom 4 of a fair matching algorithm. Hence, it must be that $M = M_0$. This completes the proof of Theorem 4.2. Q.E.D.

5. THE IMPOSSIBILITY THEOREMS

We will show in this section that not all preference profiles admit a fair matching algorithm. We will describe the structure of a certain class of preference profiles, which do.

We start by developing a matching algorithm μ_0, which, as we will see later, is fair over certain preference profiles. We need the following terminology for that purpose.

Given a preference profile p and positive integers s and t, which are at most n, then a *(s, t)-couple* with respect to p is a couple in which the man ranks the woman as his number s choice, while the woman ranks the man as her number t choice. The number of such couples will be denoted by $c(s, t; p)$.

We will define the algorithm μ_0 recursively, in the sense that we will show how it matches certain individuals in what we will call the first round of the matching. Then the procedure for the second round is the same as the procedure for the first, but applied to the fewer individuals, who were not matched by the first, after their preferences are relisted to reflect the fact that the set of available spouses is now smaller, and so on.

Before we proceed to describe the first round of the matching using μ_0, we need another definition. Let $\leq$ be the partial order defined on the set $S=\{(s, t) \mid s,t$ are positive integers $\leq n\}$ as follows: $(s_1, t_1) \leq (s_2, t_2)$, if and only if one of the following mutually exclusive conditions hold: (1) $s_1 = s_2$ and $t_1 = t_2$, (2) $\max(s_1, t_1) < \max(s_2, t_2)$, (3) $\max(s_1, t_1) = \max(s_2, t_2)$ and $\min(s_1, t_1) < \min(s_2, t_2)$. Let S_p be the subset of S, defined by $S_p = \{(s, t) \mid (s, t) \varepsilon S$ and $c(s, t; p) > 0\}$. It can easily be seen that S_p is not empty. Let (s_0, t_0) be a minimal element of S_p under the partial order $\leq$. Then, clearly $(s, t) < (s_0, t_0)$ implies $c(s, t; p) = 0$. We will call (s_0, t_0) an *initial level of reciprocity* of the preference profile p.

We now are in a position to describe the first round of the matching algorithm μ_0. Let (s_0, t_0) be the initial level of reciprocity of the preference profile p and consider the following cases:

Case 1: $s_0 = t_0$, then μ_0 matches all the (s_0, t_0)-couples.

Case 2: $s_0 \neq t_0$ and $c(t_0, s_0; p) = 0$, then μ_0 matches all the (s_0, t_0)-couples.

Case 3: $s_0 \neq t_0$, $c(t_0, s_0; p) > 0$, and $c(t_0, s_0; p) < c(s_0, t_0; p)$, then μ_0 matches all the (s_0, t_0)-couples.

Case 4: $s_0 \neq t_0$, $c(t_0, s_0; p) > 0$, and $c(s_0, t_0; p) < c(t_0, s_0; p)$, then μ_0 matches all the (t_0, s_0)-couples.

Case 5: $s_0 \neq t_0$, $c(t_0, s_0; p) > 0$, and $c(s_0, t_0; p) = c(t_0, s_0; p)$, then μ_0 cannot be applied and we will say that is is *jammed*.

A preference profile p, which jams the matching algorithm μ_0 in its first matching round will be called a *symmetric* preference profile. Otherwise, it will be called *asymmetric*.

The following theorem shows that the algorithm μ_0 does not deviate from the mutual agreement match discussed in Section 3, when such a match is possible.

THEOREM 5.1. *If p is an acyclic preference profile, then p is asymmetric and the match generated by μ_0 is the same as the mutual agreement match.*

Proof. The proof follows easily from the fact that if p is acyclic then the reciprocity level of p is (1, 1) in each round of the matching and μ_0 generates the same couples in each round as the mutual agreement match. Q.E.D.

An asymmetric preference profile for which μ_0 generates a match in one round will be called a *perfect asymmetric* preference profile.

THEOREM 5.2. *If p is a perfect asymmetric preference profile, then μ_0 is a fair matching algorithm over p.*

Proof. Let M_0 be the match generated by μ_0. To prove the theorem, we show that M_0 satisfies the axioms of a fair match. Let (s_0, t_0) be the initial level of reciprocity of p. If $s_0 \neq t_0$, assume without loss of generality that $c(s_0, t_0; p) > c(t_0, s_0; p)$. Then, because p is perfect, every couple in M_0 is a (s_0, t_0)-couple. This implies that M_0 is stable since if it were not then there exists a (s, t)-couple such that $s < s_0$ and $t < t_0$ which contradicts the definition of (s_0, t_0). Thus, Axiom 4 is proved. Axiom 3 (Maximin Optimality) follows from the definition of (s_0, t_0). Axiom 2 (Peer Indifference) is obviously satisfied, since the spouse of each individual is determined by only the preferences of that individual. Axiom 1 (Gender Indifference) is satisfied, because $c(s_0, t_0; p) > c(t_0, s_0; p)$. Q.E.D.

THEOREM 5.3. *First Impossibility Theorem. There exists an acyclic (hence asymmetric) preference profile over which there exists no fair matching algorithm.*

Proof. Suppose that the preference profile p_0 is such that man x_0 ranks woman y_0 as his number n choice and woman y_0 ranks man x_0 as her number n choice. Suppose further that every man other than x_0 is ranked as her number one choice by the woman the man ranks as his number one choice. Then, the only stable match is the match M_0, which marries x_0 to y_0 and all the other men to their respective number one choice. Obviously, the floor of M_0 is equal to n. However, M_0 is not necessarily fair, for suppose that man x_1 finds woman y_0, to be his number 2 choice and woman y_0 finds x_1 to be her number one choice. Similarly , woman y_1 ranks man x_0 as her number 2 choice and man x_0 ranks her as his number one choice. Then, the match M_1, which marries x_1 to y_0 and x_0 to y_1 and all other individuals to their number one choice has a floor equal

to 2, but is unstable. Thus, any fair algorithm must generate M_0 to satisfy the Axiom of Stability, but then in doing so, it will violate the Maximin Axiom. Hence, no fair algorithm exists over p_0. Q.E.D.

THEOREM 5.4. *Second Impossibility Theorem. If p is a symmetric preference profile, then there exists no fair matching algorithm over p.*

We prove the theorem in case of a simple example. The proof of the general case is the same.

Proof. Let $n=2$ and denote the men by John and George and the women by Mary and Linda. Since $n=2$, a preference profile is specified, once we specify the first choice of each man and woman. Let p_s be the following preference profile: John's first choice is Mary, that of George is Linda, while Mary's first choice is George and Linda's first choice is John.

Since $n=2$, there exist only two possible matches M_1 and M_2, given by $M_1=\{$(John, Mary), (George, Linda)$\}$ and $M_2=\{$(John, Linda), (George, Mary)$\}$. The achievement profiles for each match are given by $\alpha(p_s, M_1)=(1, 1; 2, 2)$ and $\alpha(p_s, M_2)=(2, 2; 1, 1)$, where the men are listed in the order of John and then George, while the women are listed in the order of Mary and then Linda.

It is clear that p_s is symmetric, since the initial level of reciprocity is (1, 2) and $c(1, 2: p_s)=c(2, 1: p_s)=2$. It is also clear that M_1 and M_2 are both stable and are maximal elements of the partial order $j(p_s)$. Hence, by Theorem 4.2, if there exists a fair matching algorithm μ over p_s, it has to generate M_1 or M_2. But then such an algorithm will not satisfy Axiom 1 (Peer Indifference), since, if μ generates M_1 when the women are occupying File 1, then it will generate M_2 when the men are occupying File 2. Therefore, there exists no fair matching algorithm Q.E.D.

We end this section with some remarks on the results derived in it.

Note that in the example in the proof of the first impossibility theorem, if it were not for the presence of the individuals x_0 and y_0, there exists a fair matching algorithm, this being simply the one that matches each individual to his (resp. her) number one choice. Thus, the example shows that the inclusion of new individuals into the group to be matched can have a destabilizing influence (literally) on the possibility of a fair match,

depending on the preferences of the new individuals. This may lead to the possibility of some individuals wanting to exclude others from the community of individuals to be matched, simply because the fairness axioms will result in a decrease in the achievement of these individuals in the match. In other words, the requirements of fairness can lead to prejudice.

The matching algorithm μ_0 is in a sense the generalization of the mutual agreement match. In case the preference profile p is acyclic, then μ_0 applied to p results in the mutual agreement match.

6. STABLY FAIR ALGORITHMS

We start by considering the implications of the impossibility theorems. The example in the proof of the first impossibility theorem shows that even the algorithm which generates the mutual agreement match (we will refer to this algorithm as the *mutual agreement algorithm*) has no chance to be fair because of conflict between stability and maximin optimality. Most readers may consider this to be an untenable situation, since the mutual agreement match is the outcome of free choice and, hence, one would like it to be fair. Thus, an objection may be raised that the reason that the first impossibility theorem is true is the fact that we are requiring the maximin optimal match to be stable, which is too strong a requirement, and that, if we were to restrict ourselves to stable matches in the statement of the axioms of a fair matching algorithm, the outcome might be different. We will consider in this section such a relaxation of the definition of fairness. However, it should be stressed at this point that in doing so we are giving preference to stability over maximin optimality. This implies that we are favoring the freedom of choice that stability implies (freedom to divorce) to the welfare of the worst-off individual as reflected in the maximin principle, while in our approach so far stability and the maximin principle were on an equal footing. It is this favoritism that will enable us eventually to conclude that the mutual agreement algorithm is fair. If we reverse our approach and drop stability and hence prohibit divorce, then the mutual agreement match is unfair and hence the individuals have to be disallowed a free choice in selecting their partners in case of acyclic preferences. Thus, a prohibition of the freedom to dissolve a marriage implies a prohibition of the freedom to enter a

marriage. We proceed now to state the formal definition of a stably fair algorithm.

Let $p \in \Pi(X, Y)$ and let $\mathscr{M}_s(p)$ denote the set of stable matches with respect to p. This set is not empty by Theorem 1 of Gale and Shapley (1962). A matching algorithm over $\Lambda \subseteq \Pi(X, Y)$ is said to be *stably fair*, if it satisfies Axioms 1 through 3 in Section 4, with $\mathscr{M}_s(p)$ replacing $\mathscr{M}$ in the statement of these axioms. Obviously, a fair algorithm is stably fair, but the converse is not true as we shall see later. Thus, stable fairness is a weaker concept than fairness.

Let Λ_{ac} be the set of acyclic preference profiles. Since for every $p \in \Lambda_{ac}$ the set $\mathscr{M}_s(p)$ consists of one element, which is the mutual agreement match, we have the following

THEOREM 6.1. *The mutual agreement algorithm is a stably fair, but not fair, algorithm over the set of acyclic preference profiles.*

The reason that the mutual agreement algorithm is not fair is the example in the proof of the first impossibility theorem.

We now consider the question whether μ_0 is stably fair over the set of asymmetric preferences.

THEOREM 6.2. *The matching algorithm μ_0 is not stable over every asymmetric preference profile.*

Proof. Assume $n > 3$. Let p be an asymmetric preference profile, such that its initial level of reciprocity is (2, 3). Suppose that $c(3, 2: p) < c(2, 3; p) = n - 1$. Thus, every man except for one, say x_0, can be matched in a match M_0 generated by μ_0 to a woman whom he ranks as his number 2 choice and she ranks him as her number 3 choice. Let the woman with whom x_0 has to be matched in M_0 be y_0. Suppose x_0 ranks y_0 as his number n choice and y_0 ranks x_0 as her number n choice.

Suppose now that there exists a man x_1 who ranks woman y_0 as his number 1 choice. Then x_1 and y_0 prefer each other to their present spouses. Thus M_0 is unstable. Q.E.D.

Since μ_0 is not even stable over asymmetric preference profiles, obviously then μ_0 is not stably fair.

The above proof shows again how the presence of a particular individu-

al (woman y_0) can destabilize an otherwise fair match. This shows that the stability requirement is quite a strong requirement and that, if more requirements are added to it, then accomplishing fairness in the matching becomes generally impossible.

The following theorem is a restatement of the second impossibility theorem for stably fair algorithms.

THEOREM 6.3. *If p is a symmetric preference profile, then there exists no stably fair matching algorithm.*

It is easy to see that the same argument given in the proof of Theorem 5.4 yields the above theorem.

7. THE ALGORITHM μ_1

Let us summarize our results so far. We started by stating a set of axioms which define a fair algorithm and we showed why we consider these axioms a 'natural' definition of a fair algorithm, in the sense that all rational individuals involved in the match will agree that these axioms serve their best interests in generating a fair match. Then, we defined Rawls' principle of justice j through a partial order and showed that Rawls' principle of justice is compatible with the concept of a fair algorithm, in the sense that a fair matching algorithm has to generate a match, which is a maximal element of j.

Given this concept of fairness, we showed that there are preference profiles over which a fair matching algorithm exists and others for which there exists no such algorithm. Because the mutual agreement match can turn out to be unfair, we decided that our concept of fairness is 'too strong' in a sense and we weakened it by defining stably fair algorithms.

In case of stable fairness we found again that some preference profiles admit a fair matching algorithm, while others still do not; but the mutual agreement match is a stably fair match.

What we would like to do now is to find out, if some kind of a testing procedure can be developed such that given a preference profile p, then it is possible to determine whether a stably fair matching algorithm exists over p or not and in case there exists such an algorithm, construct the fair match generated by it over p.

In order to accomplish the above stated goal, we return to the algorithm μ_0. We will show how μ_0 can be adjusted to yield a stably fair matching algorithm μ_1 over a certain class of preference profiles.

Let p be an asymmetric preference profile. The first round of the matching by μ_1 over p is identical to that of μ_0.

Let A_1 and B_1 be sets of men and women, respectively, matched by μ_0 in the first round. If $A_1 = X$ and $B_1 = Y$, then μ_1 agrees with μ_0 and p must be perfect. We have already shown (Theorem 5.2) that in this case a fair match is generated. Thus, in particular it is also stably fair.

If $B_1 \neq \emptyset$, we proceed to describe how the first round for μ_1 adjusts the first round of μ_0 so that the resulting match is stable. We define two sets B_1' and B_1'', whose union is Y, and two sets A_1' and A_1'', whose union is X, by the following decision rule, denoted by d.

Decision Rule d

Step 1: Let y be a woman in $Y\text{-}B_1$. Let $A_1(y)$ be the set of men in A_1 which y prefers to every man in $X\text{-}A_1$. If $A_1(y) = \emptyset$, then d places y in B_1''. If $A_1(y)$ is not empty, let $\hat{A}(y)$ be the set of men in $A_1(y)$ who prefer y to their present spouses generated by μ_0. If $\hat{A}_1(y)$ is empty, then y is placed on B_1''. If $\hat{A}(y)$ is not empty, let x be the man y prefers the most in $\hat{A}_1(y)$ and let y_1 be the current spouse in B_1 as generated by μ_0. Then, y is placed in B_1' as the spouse of x and y_1 is placed in B_1''.

Step 2: Let y be a woman in B_1. If Step 1 of d did not place y in B_1'', then y is placed in B_1', this remaining the spouse of a man in A_1 as generated by μ_0.

Step 3: Repeat Steps 1 and 2 applied to the *men* with the set B_1' playing the role of A_1 and A_1 playing the role of B_1 in Steps 1 and 2, and thus generating the sets A_1' and A_1''.

One can easily check that the dual statement of decision rule d which is obtained from the above statement by replacing the word woman by man and the word man by woman and references to sets of individuals of either sex adjusted accordingly generates the same pairs of sets A_1', A_1'' and B_1', B_1'' as the first statement. Thus, d is gender-indifferent.

Obviously, it may happen that after the application of d we may have $A_1' = A_1$ and $B_1' = B_1$; but, in general we will end up with two new sets replacing A_1 and B_1. The set B_1' will be the set of the spouses of the men

in A_1' after d is applied. This completes the description of the first round of μ_1.

The second round of μ_1 consists of the first round applied to the set of men consisting of $A_1'' = X\text{-}A_1'$ and the set of women consisting of $B_1'' = Y\text{-}B_1'$, and so on.

If in any one of the rounds of μ_1 we are unable to apply μ_1 because the preference profile for that round is symmetric, then we will call the original preference profile *inadmissible.* If by applying μ_1 round after round we end up matching all the men and all the women, then the preference profile p will be called *admissible.*

THEOREM 7.1. *The Fairness Test. There exists a stably fair matching algorithm over a preference profile p, if and only if p is an admissible preference profile.*

If p is an admissible preference profile, then there exists a unique stably fair matching algorithm over p, which is μ_1.

Proof. To prove the first statement of the theorem we note that if p is inadmissible then in one of the rounds of μ_1 applied to p we will end up trying to match one group of women B to one group of men A whose preferences are symmetric and thus by the Second Impossibility Theorem (6.3) no stably fair algorithm exists to match these two groups. Therefore, any stably fair matching algorithm has to match eventually B to $X\text{-}A$ and A to $Y\text{-}B$. But, because of the construction of μ_1 such an algorithm will be unstable.

To prove the second statement of the theorem, we first note that μ_1 is a stably fair matching algorithm. To see this, we first note that decision rule d insures that the match generated by μ_1 is stable, and the fact that μ_1 agrees with μ_0 in picking the couples which are allowed to marry if their match is stable insures that there exists no other stable match than the one generated by μ_1 which has a lower floor. Thus, μ_1 satisfies the Maximin Axiom. Peer Indifference is guaranteed, since the men and the women are matched by μ_1 according to their preferences and not the order in which they are listed. It remains to check that μ_1 satisfies the Gender Indifference Axiom. Since μ_0 does not allow gender favoritism, the only possibility that μ_1 does allow gender favoritism is through the decision rule d. But we have already noted that the decision rule d is gender-indifferent.

It remains to show that μ_1 is unique. But this follows easily from the fact that the construction of μ_1 shows that when p is an admissible preference profile the set of the maximal elements of the partial order $j(p)$ over $\mathscr{M}_s(p)$ has only one element. Hence, by Theorem 4.2, any stably fair algorithm over p has to generate the same match as μ_1. Q.E.D.

Knuth (1976) reports that Selkow has developed an algorithm which Knuth claims to be fair. However, his description of the algorithm is very sketchy and ambiguous and the algorithm seems not to satisfy gender indifference, since at a certain stage individuals are chosen at random.

REFERENCES

Gale, D. and L. S. Shapley: 1962, 'College Admission and the Stability of Marriage', *American Mathematical Montly* **69**, 9–14.

Knuth, D.: 1976, *Mariages Stables*, Montréal: Les Presses de l'Université de Montréal.

Masarani, F. and S. S. Gokturk: 1986, 'A Pareto Optimal Characterization of Rawls' Social Choice Mechanism', *Journal of Mathematical Economics* **15**, 157–170.

Rawls, J.: 1971, *A Theory of Justice*, Cambridge, Mass.: Harvard University Press.

Rochford, S. C.: 1984, 'Symmetrically Pairwise-Bargained Allocations in an Assignment Market', *Journal of Economic Theory* **34**, 262–281.

Roth, A. E.: 1985a, 'Conflict and Coincidence of Interest in Job Matching: Some New Results and Open Questions', *Mathematics of Operations Research* **10**, 379–389.

Roth, A. E.: 1985b, 'Common and Conflicting Interests in Two-Sided Matching Problems', *European Economic Review* **27**, 75–96.

Department of Mathematics,
Hostos College of C.U.N.Y.
Bronx, N.Y. 10451, U.S.A.

and

Department of Economics,
St. John's University,
Jamaica, N.Y. 11439, U.S.A.